08/08/24

MAP

MLE
(Maximum likelihood estimate).

n data points.

$D = \{x_1, x_2, \ldots x_n\}$

likelihood of these data points, given a model $\theta = \{ \quad \}$

$$\log \hat{P}(D|\theta) = \hat{P}(\{x_1, x_2, \ldots x_n\}|\theta) = \boxed{\prod_{k=1}^{n} \hat{p}(x_k|\theta)}$$

the MLE of θ is : $\theta_{MLE} = \arg\max_{\theta} \hat{p}(D|\theta)$

Coin flip
H | T

$$\boxed{\theta = P(H) = q = \frac{\# \text{ of } H}{\text{total Count}} = \frac{n_1}{n_1 + n_2}}$$

$D = \{x_1, x_2 \ldots x_n\}$

$n_1 : H$

$n_2 : T$

$$P(D|\theta) = \prod_{i=1}^{n} P(x_i|\theta) = q^{n_1}(1-q)^{n_2}$$

To obtain $\hat{q} = \arg\max_{q} q^{n_1}(1-q)^{n_2} = F$

$$\frac{\partial F}{\partial q} = n_1 q^{n_1 - 1}(1-q)^{n_2} - q^{n_1}(1-q)^{n_2 - 1} \times n_2 = 0$$

$$\Rightarrow q^{n_1 - 1}(1-q)^{n_2 - 1}(n_1(1-q) - q n_2) = 0$$

$$\Rightarrow q = \frac{n_1}{n_1 + n_2}$$

$$\log \hat{P}(D|\theta) = \sum_{i=1}^{N} \log(x_i|\theta) \qquad D = \{x_1 \ldots x_n\}$$

MAP : Maximum a posteriori Estimate

$$\log(\theta_{MAP}) = \arg\max_{\theta} \underbrace{P(\theta|D)}_{\text{Posterior prob}} = \arg\max_{\theta} \underbrace{P(D|\theta)}_{\text{likelihood}} \underbrace{P(\theta)}_{\text{prior}}$$

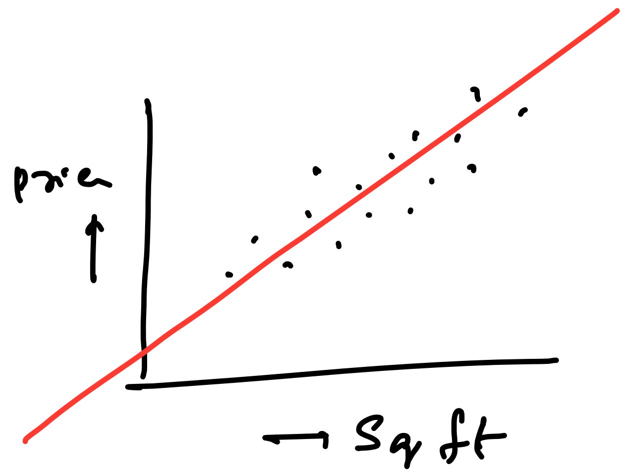
$$= \log P(\theta) + \log P(D|\theta)$$

$$= \log P(\theta) + \sum_{i=1}^{n} P(x_i|\theta)$$

Linear Regression

Living area (Sqft)	price ($)
2104	400
1600	330
3000	232
→ 2500	?

$|f(x) - y|$



$h_\theta(x) = \theta_0 + \theta_1 x_1 + \theta_2 x_2 + \dots$

$= \theta_0 + \sum_{i=1}^{m} \theta_i x_i$

x_0

$= \sum_{i=0}^{m} \theta_i x_i$

$x_0 = 1$

minimize the distance from all the points to the line L

$f(x_1, x_2, x_3) = x_1 + 3x_2 + 4x_3$

$(1, 2, 3)$

$f = \theta_1 x_1 + \theta_2 x_2 + \theta_3 x_3 =$

Loss: $\frac{1}{2}\sum_{i=1}^{m} (f(x_i) - y_i)^2$

$\theta = \{\theta_0, \theta_1, \dots, \theta_n\}$

$d_i = (x_i, y_i)$

d_1
d_2
d_3
$\vdots$
d_m

$$J(\theta) = \sum_{i=1}^{m} (h_\theta(x^i) - y^i)^2$$

m = # of training instances.

Superscript → instance

j = Subscript → = parameter / feature

$$\frac{\partial J}{\partial \theta_j} = \frac{\partial}{\partial \theta_j} \frac{1}{2}(h_\theta(x^i) - y^i)^2$$

$$= 2 \times \frac{1}{2} (h_\theta(x^i) - y^i) \frac{\partial}{\partial \theta_j} (h_\theta(x) - y)$$

$$= (h_\theta(x^i) - y^i) \times \frac{\partial}{\partial \theta_j} \left(\sum_{k=1}^{n} \theta_k x_k - y\right)$$

$$= (h_\theta(x^i) - y^i)\, x_j$$

old

$\hat{\theta}_j \leftarrow \theta_j - \eta \frac{\partial J}{\partial \theta_j}$

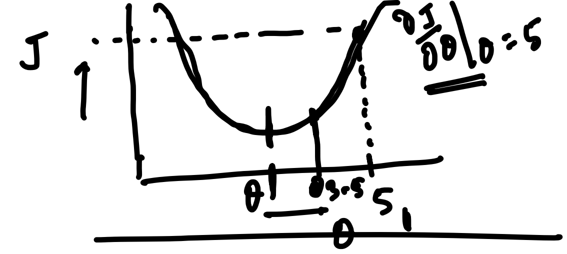
new

learning rate

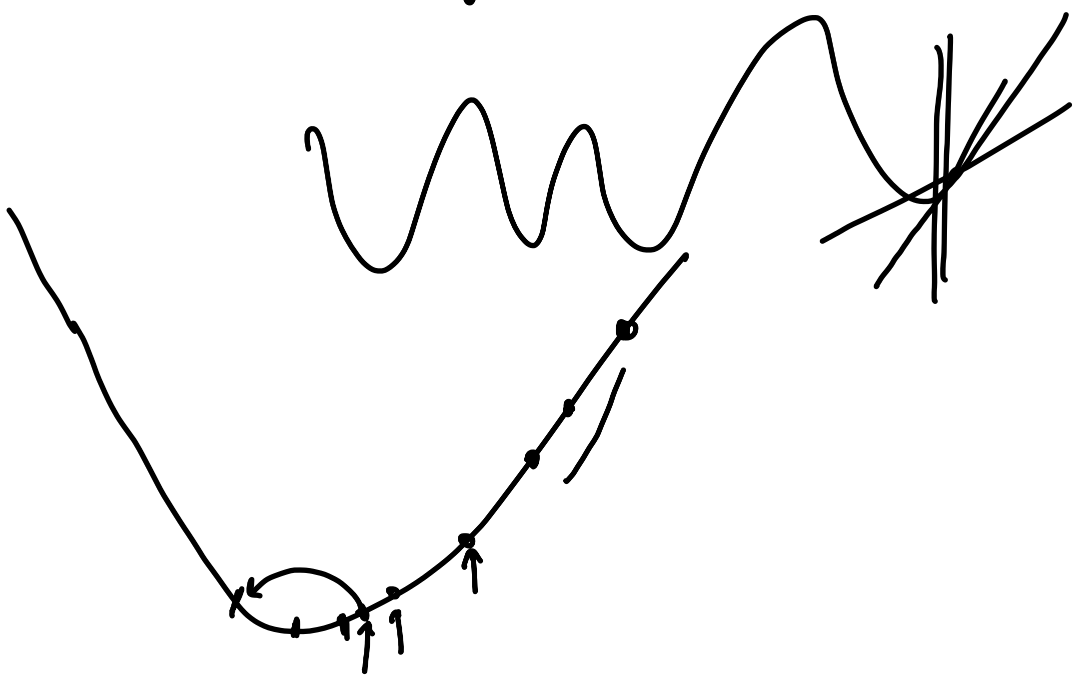
$$\frac{\partial J}{\partial \theta_p} = (h_\theta(x^i) - y^i)\, x_p$$

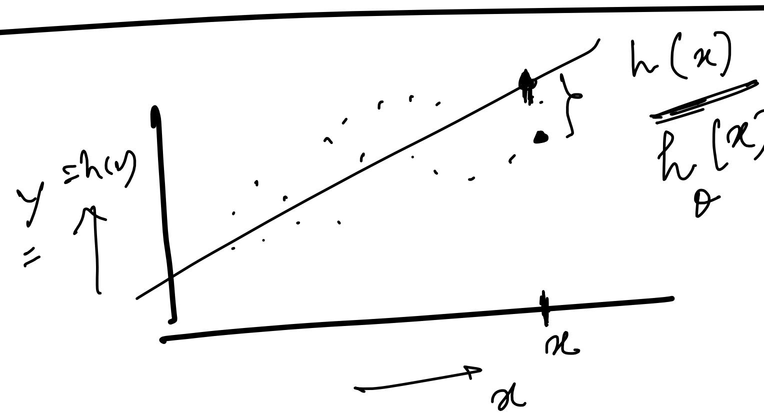
Batch GD | Stochastic GD | mini Batch GD

$$\hat{\theta}_j \leftarrow \theta_j - \eta \frac{\partial J}{\partial \theta_j}$$



$\theta = 5$





22/08/24

$$y = w_0 + w_1 x_1 + w_2 x_2 \cdots$$

θ

m = # of data points

n = # of features

$$\frac{1}{2} \sum_{i=1}^{m} \left(h_\theta(x^i) - y^i \right)^2$$

$$x = \{ \; ---- \; \}_n$$

The Normal Equation

$$f : \mathbb{R}^{m \times n} \rightarrow \mathbb{R}$$

$$f(A) = \frac{3}{2} A_{11} + 5 A_{12}^2 + A_{21} A_{22}$$

$$A = \begin{bmatrix} A_{11} & A_{12} \\ A_{21} & A_{22} \end{bmatrix} \qquad \nabla_A f(A) = \begin{bmatrix} \frac{\partial f}{\partial A_{11}} & \frac{\partial f}{\partial A_{12}} \\ \frac{\partial f}{\partial A_{21}} & \frac{\partial f}{\partial A_{22}} \end{bmatrix}$$

$$= \begin{bmatrix} 3/2 & 10 A_{12} \\ A_{22} & A_{21} \end{bmatrix}$$

$$\text{tr}(A) = \text{tr} A = \sum_{i=1}^{n} A_{ii}$$

1. $\text{tr}(ABC) = \text{tr}(CAB) = \text{tr}(BCA)$
2. $\text{tr}(A) = \text{tr}(A^T)$
3. $\text{tr}(A+B) = \text{tr}(A) + \text{tr}(B)$
4. $\text{tr}(aA) = a \, \text{tr} A$

(5) $\nabla_A \operatorname{tr}(AB) = B^T$ (6) $\nabla_{A^T} f(A) = (\nabla_A f(A))^T$

(7) $\nabla_A \operatorname{tr} ABA^TC = CAB + C^TAB^T$

$$\vec{X} = \begin{bmatrix} — x^{(1)T} — \\ — x^{(2)T} — \\ — x^{(3)T} — \\ \vdots \\ — x^{(m)T} — \end{bmatrix}_{m \times n} \qquad \vec{Y} = \begin{bmatrix} y^1 \\ y^2 \\ \vdots \\ y^m \end{bmatrix}_{m \times 1} \qquad \vec{\theta} = \begin{bmatrix} \\ \\ \\ \end{bmatrix}_{n \times 1}$$

$$(\vec{X}\vec{\theta} - \vec{Y}) = \begin{bmatrix} h_\theta(x^1) - y^1 \\ h_\theta(x^2) - y^2 \\ \vdots \\ h_\theta(x^m) - y^m \end{bmatrix} =$$

$$J(\theta) = \frac{1}{2}\sum \left(h_\theta(x^i) - y^i\right)^2 = \frac{1}{2}(X\theta - Y)^T(X\theta - Y)$$

$$\nabla_\theta J(\theta) = \frac{1}{2}\nabla_\theta (X\theta - Y)^T(X\theta - Y)$$

$$= \frac{1}{2}\nabla_\theta (\theta^T X^T - Y^T)(X\theta - Y)$$

$$= \frac{1}{2}\nabla_\theta (\theta^T X^T X\theta - \theta^T X^T Y - Y^T X\theta + Y^T Y)$$

$$= \frac{1}{2}\left[\nabla_\theta \operatorname{tr}(\theta^T X^T X\theta - \theta^T X^T Y - Y^T X\theta)\right]$$

$$= \frac{1}{2}\left[\nabla_\theta \operatorname{tr}(\theta^T X^T X\theta) - \operatorname{tr}(\theta^T X^T Y) - \operatorname{tr}(Y^T X\theta)\right]$$

A^T A

$$= \frac{1}{2}\left[\nabla_\theta \operatorname{tr}(\theta^T X^T X\theta) - 2\nabla_\theta \operatorname{tr}(Y^T X\theta)\right]$$

$A = Y^T X\theta$

$A^T = (Y^T X\theta)^T$

$= \theta^T X^T Y$

$$= \frac{1}{2}\left(\nabla_\theta \operatorname{tr}(\theta^T X^T X\theta) - 2X^T Y\right)$$

$\nabla_\theta \operatorname{tr}(Y^T X\theta)$

$= \nabla_\theta \operatorname{tr}(\theta \underbrace{Y^T X}_{B})$

$= (Y^T X)^T$

$\nabla_A \operatorname{tr}(AB) = B^T$

$\nabla_\theta \operatorname{tr}(\underbrace{\theta^T}_{A} \underbrace{X^T X}_{B} \underbrace{\theta}_{A^T})$ $C = I$

$\cancel{\theta} = A^T$

$\nabla_\theta \operatorname{tr}(\underbrace{\theta}_{A} \underbrace{\theta^T}_{B}\ \underbrace{X^T X}_{C})$ A^T $= I$

$\nabla_A \operatorname{tr} ABA^TC$

$= CAB + C^TAB^T$

$$= X^T X\theta + X^T X\theta$$

$$= 2X^T X\theta$$

$\Rightarrow \frac{1}{2}(2X^T X\theta - 2X^T Y) = 0$

$\boxed{X^T X\theta = X^T Y}$ $\longrightarrow$ Normal Eq

$\theta = (X^T X)^{-1} X^T Y$ if $X^T X$ is invertible

$y = w_0 + w_1 x_1 + \dots + w_n x_n$

$= w_0 + w_1 x_1^2 + w_2 x_2^2 + \dots + w_n x_n^2 \leftarrow$

linear in the parameters.

$Y = w_0 + w_1 \phi(x_1) + w_2 \phi(x_2) + \dots + w_n \phi(x_n)$

ϕ = non-linear basis funct.

$\phi_j(x) = x^j$ – polynomial

$= \frac{(x-\mu)^2}{2\sigma^2}$ – Gaussian

$= \frac{1}{1+\exp(-s_i x_i)}$ – Sigmoid

$$J = \sum_{i=1}^{m}\left(y^i - \sum_{j=1}^{n}\theta_j \phi(x^i)\right)^2 = \sum\left(y^i - \theta^T \underbrace{\phi(x^i)}_{x^i}\right)^2$$

$$\frac{\partial J}{\partial \theta} = \sum (y^i - \theta^T x^i)^2 = 2\sum (y^i - \theta^T x^i)\, x^{(i)} = 0$$

$$\sum y^{(i)} x^{(i)T} = \theta^T \sum x^{(i)T} x^{(i)}$$

$\Rightarrow X^T Y = (X^T X)\theta$ — Normal

$\Rightarrow \boxed{\theta = (X^T X)^{-1}(X^T Y)}$

$X X^T$ $(X^T X)$ = non-invertible

$X \mid m \times n$

Regularizer

Ridge

$$\hat{\theta} = \arg\min \sum_{i=1}^{m} (y_i - x_i\theta)^2 + \lambda \|\theta\|_2^2$$

L2 Ridge Regression

$$\hat{\theta} = (X^TX + \lambda I)^{-1} X^T Y$$

L1 Lasso Regression

$$\hat{\theta} = \sum_{i=1}^{m} (y_i - x_i\theta)^2 + \lambda \|\theta\|_1$$

Prove that : $\nabla_A \operatorname{tr}(ABA^TC) = C^TAB^T + CAB$

→ we use two theorems as follows:

$\operatorname{tr}(ABCD) = \operatorname{tr}(DABC) = \operatorname{tr}(CDAB) = \operatorname{tr}(BCDA)$ — (I)

chain Rule of matrix differentiation:

$d(AB) = d(A)B + A\,d(B)$ — (II)

Therefore, $\nabla_A \operatorname{tr}(ABA^TC) = \nabla_X \operatorname{tr}(XBA^TC) + \nabla_X \operatorname{tr}(ABX^TC)$

$\Rightarrow \nabla_X \operatorname{tr}(XBA^TC) = (BA^TC)^T \quad (\because \nabla_A \operatorname{tr}(AB) = B^T)$

$= C^TAB^T$ — (III)

$\nabla_X \operatorname{tr}(ABX^TC) = \nabla_X \operatorname{tr}(X^TCAB)$ (due to (I))

$\Rightarrow ((\nabla_X \operatorname{tr}(X^TCAB)^T)^T$

as $\nabla_{A^T} f(A) = (\nabla_A f(A))^T \Rightarrow (\nabla_{X^T} \operatorname{tr}(X^TCAB))^T$

as $\nabla_A \operatorname{tr}(AB) = B^T \Rightarrow ((CAB)^T)^T$

$\Rightarrow CAB$ — (IV)

There from (III) and (IV)

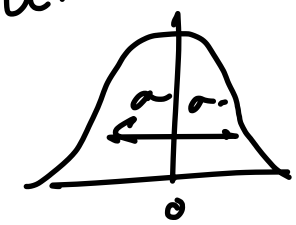
$\nabla_A \operatorname{tr}(ABA^TC) = C^TAB^T + CAB$

Probabilistic Interpretation of Linear Reg.

29/08/24

$$y_i^i \neq \theta^T x_i^i = h_\theta(x_i^i)$$

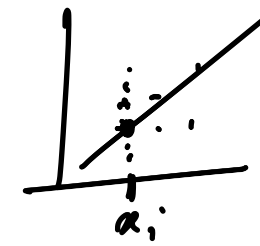
$$y^i = \theta^T x^i + \epsilon^i$$ error ⇒ unmodelled error



$E(\epsilon^i) = 0$

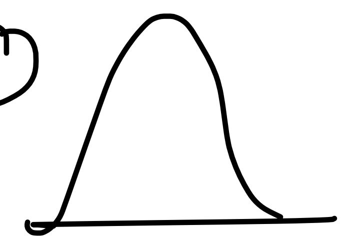
σ^2

$E(y^i) \approx E(\theta^T x^i)$



$$p(\epsilon^i) = \frac{1}{\sqrt{2\pi}\,\sigma} \exp\left(-\frac{(\epsilon^i)^2}{2\sigma^2}\right)$$

$$p(y^i \mid x^i; \theta) = \frac{1}{\sqrt{2\pi}\,\sigma} \exp\left(-\frac{(y^i - \theta^T x^i)^2}{2\sigma^2}\right) \quad (1) =$$



$$y^i \mid x^i; \theta \sim \mathcal{N}(\theta^T x^i, \sigma^2)$$

$\epsilon \sim$ iid

$$L(\theta) = L(\theta; \vec{X}, \vec{Y}) = p(\vec{y} \mid \vec{X}; \theta) \leftarrow$$

design matrix

vector of labels

from (1)

$$= \prod_{i=1}^{m} p(y^i \mid x^i; \theta) = \prod_{i=1}^{m} \frac{1}{\sqrt{2\pi}\,\sigma} \exp\left(-\frac{(y^i - \theta^T x^i)^2}{2\sigma^2}\right)$$

maximize $L(\theta) \equiv \max \log L(\theta) = \ell(\theta)$

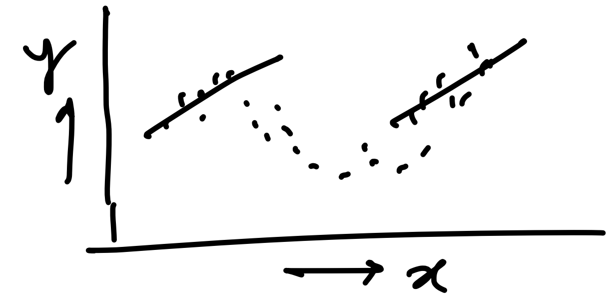
$$\ell(\theta) = \log L(\theta)$$

$$= \sum_{i=1}^{m} \log\left(\frac{1}{\sqrt{2\pi}\,\sigma} \exp\left(-\frac{(y^i - \theta^T x^i)^2}{2\sigma^2}\right)\right)$$

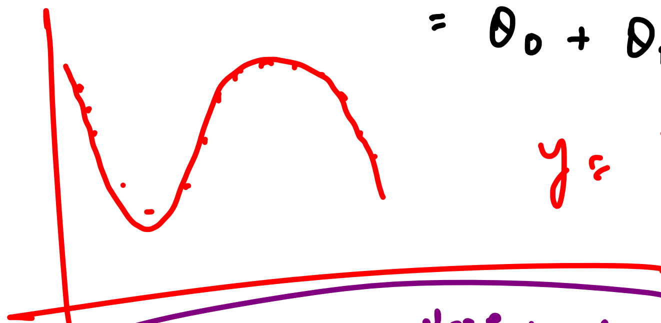
$$= m \log \frac{1}{\sqrt{2\pi}\,\sigma} - \frac{1}{\sigma^2} \sum_{i=1}^{m} \frac{(y^i - \theta^T x^i)^2}{2} \leftarrow ?$$

independent of θ

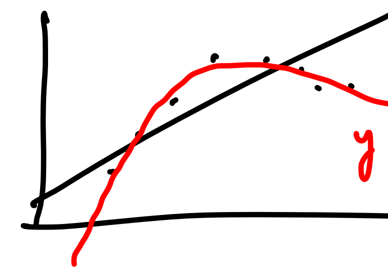
Locally weighted LR



$$y = \theta_0 + \theta_1 x$$
$$= \theta_0 + \theta_1 x + \theta_2 x^2$$
$$= \theta_0 + \theta_1 x + \theta_2 x^{1/2}$$

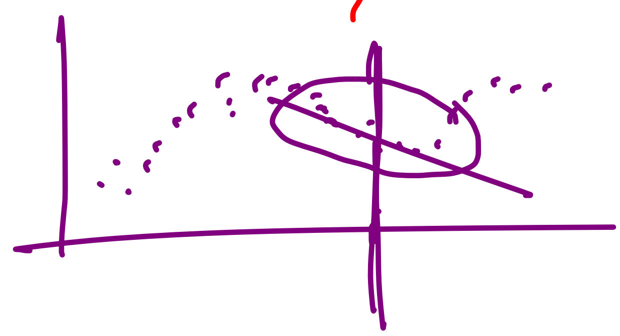


$$y = \sum_{i=1}^{5} \theta_i x^i$$



$y = \theta_0 + \theta_1 x + \theta_2 x^2$

Non-parametric model



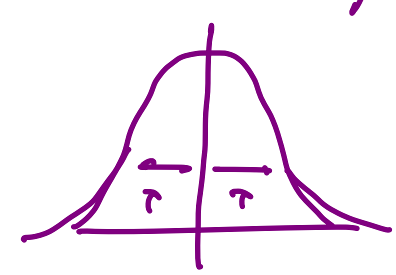
Fit θ by minimizing

$$\sum_{i=1}^{m} w^i \left(y^i - \theta^T x^i \right)^2 \qquad w^i = \exp\left(- \frac{(x^i - x)^2}{2\tau^2} \right) \in$$

τ = bandwidth

= broad → τ = over smooth thing / underfit

= narrow → jagged fit / over fit



SSE = SST − SSM → Board.

Logistic 'Regression'

31/08/24

- Classification method → binary

$y \to 0, 1 \;/\; -1, 1$

$0/-1 \to$ neg.

$1/+1 \to$ positive.

$x_i \to \{x_1, x_2, \ldots, x_n\}$ $\qquad x_0 = 1$

$$h_\theta(x) = \theta^T X = \theta_0 + \sum_{i=1}^{n} \theta_i x_i = \sum_{i=0}^{n} \theta_i x_i$$

$p(+)$ $\quad$ $p(-)$

$$h_\theta(x) = g(\theta^T x) = \frac{1}{1+e^{-\theta^T x}} = \alpha(\theta^T x)$$

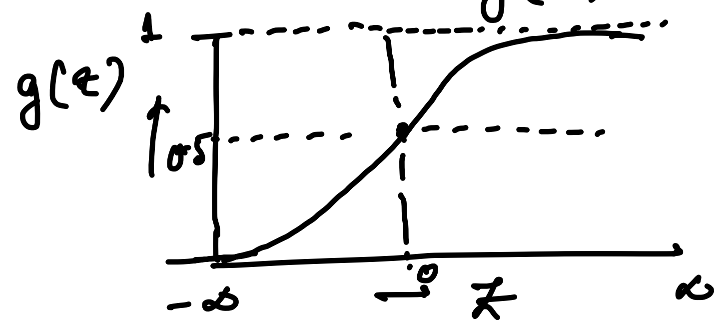
$$\boxed{g(z) = \frac{1}{1+e^{-z}}}$$ → logistic function ← Sigmoid fn

$z = 0$

$g = \frac{1}{2}$

$g(z) \to 1 ; z \to \alpha$

$g(z) \to 0 ; z \to -\infty$



$$\frac{\partial g}{\partial z} = \frac{\partial}{\partial z} \frac{1}{1+e^{-z}}$$

$$= \frac{1}{(1+e^{-z})^2}(e^{-z})$$

$$= \frac{1}{(1+e^{-z})}\left(\frac{e^{-z}}{1+e^{-z}}\right) = \frac{1}{1+e^{-z}}\left(1 - \frac{1}{1+e^{-z}}\right)$$

$$= g(z)(1-g(z))$$

$$g'(z) = g(z)(1-g(z))$$

Assume that $p(y=1 \mid x^i; \theta) = h_\theta(x^i)$

$(y=0 \mid x^i; \theta) = 1 - h_\theta(x^i)$

$x_i \mid i = 1, \dots, m$

$$p(\vec{y} \mid \vec{x}; \theta) = \prod_{i=1}^{m} p(y=1 \mid x^i) \dots p(y=0 \mid x^i)$$

$$= \prod_{i=1}^{m} \left(h_\theta(x^i)\right)^{y^i} \left(1 - h_\theta(x^i)\right)^{1-y^i}$$

$$L(\theta) = p(\vec{y} \mid \vec{x}; \theta) = \prod_{i=1}^{m} h_\theta(x^i)^{y^i} \left(1 - h_\theta(x^i)\right)^{1-y^i}$$

maximize $L(\theta)$

$$\ell(\theta) = \log L(\theta) = \sum_{i=1}^{m} y^i \log h_\theta(x^i) + (1-y^i)\log(1-h_\theta(x^i)) \quad \text{--- (1)}$$

$$\theta \leftarrow \theta + \nabla_\theta \ell(\theta) \text{ — Gradient ascent}$$

derivative of $\ell(\theta)$ w.r.t. $\theta = \{\theta_0, \theta_1, \theta_2, \ldots, \theta_n\}$

$$\ell(\theta) = y \log h_\theta(x) + (1-y)\log(1-h_\theta)$$

$$" g(\theta^T x)$$

$$\frac{\partial \ell(\theta)}{\partial \theta_j} = y \frac{1}{g(\theta^T x)} g'(\theta^T x) \cdot \frac{\partial}{\partial \theta_j}(\theta^T x) - (1-y)\frac{1}{1-g(\theta^T x)} g'(\theta^T x) \frac{\partial}{\partial \theta_j}(\theta^T x)$$

$$= \left(y \frac{1}{g(\cdot)} - (1-y)\frac{1}{1-g(\cdot)} \right) g'(\cdot) \frac{\partial}{\partial \theta_j}(\theta^T x)$$

$$= \left(\frac{y}{g(\cdot)} - \frac{(1-y)}{1-g(\cdot)} \right) g(1-g)\, x_j$$

$$= [y(1-g(\cdot)) - (1-y)g(\cdot)]\, x_j$$

$$= (y - g(\cdot))x_j = (y - h_\theta(x))\, x_j$$

$$\theta_j \leftarrow \theta_j + \alpha(y^i - h_\theta(x^i))\, x^i_j \quad \text{for } x_i$$

why sigmoid

$$p(y=1|x) = \sum \theta_i x_i$$

$0, 1$ $\qquad -\infty \quad \infty$

$$\text{odds} \quad \left| \frac{p(y=1|x)}{1-p(y=1|x)} \right| = \sum \theta_i x^i$$

$-\infty, \infty$

$0 - \infty$

$$\ln\left(\frac{p}{1-p}\right) = \sum \theta_i x^i$$

$$\Rightarrow \frac{p}{1-p} = e^{\Sigma(\,)}$$

$$\Rightarrow p = (1-p)e^{\Sigma}$$

we ~~prov~~ proved on whiteboard that (1) is concave

$$\Rightarrow \quad p = \frac{e^{\Sigma}}{1+e^{\Sigma}}$$

$$p(y=1|x) = \frac{e^{\Sigma \theta_i x_i}}{1+e^{\Sigma \theta_i x_i}}$$

$$p(y=0|x) = \frac{1}{1+e^{\Sigma}}$$